CLAIMS

What is claimed is:

1. A document automatic classification system, comprising:

list generation means for generating a word list for each category by extracting words from a learning document set; and

unnecessary word determination means for relatively determining an unnecessary word for each category on the basis of a frequency of appearance of a given word in each category by using the list generated by said list generation means.

2. The system according to Claim 1, wherein said list generation means generates a list indicating a frequency of appearance of a given word for each category from said learning document set in the storage means.

3. The system according to Claim 1, wherein said unnecessary word determination means extracts a word belonging to a given category and determines it to be an unnecessary word if the word appears more frequently than a given standard in another category.

4. The system according to Claim 1, wherein said unnecessary word determination means determines the word extracted from said given category to be an unnecessary word if it appears more frequently in another category than the given standard determined according to a predetermined threshold and the number of documents belonging to said another category.

5. The system according to Claim 1, further comprising:

classification catalog storage means for storing a list for each category from which unnecessary words were eliminated based on the determination with said unnecessary word determination means; and

document classification means for performing classification processing for classification target documents by using said classification catalog stored in the classification catalog storage means.

6. A document automatic classification system, comprising:

a classified document set storage device for storing documents classified according to category;

a category table generation unit for generating a table broken down by category including information on a frequency of appearance of a word contained in a document acquired from said classified document set storage device;

an unnecessary word elimination unit for eliminating an unnecessary word for each category concerned from the table on the basis of a frequency of appearance in each category of a given word acquired from the table broken down by category generated by said category table generation unit; and

a classification catalog storage device for storing the table from which the unnecessary word was eliminated by said unnecessary word elimination unit.

7. The system according to Claim 6, further comprising:

a classification target document storage device for storing classification target documents to be classified; and

a document classification processing unit for performing classification processing for the classification target documents stored in said classification target document storage device by using said table stored in said classification catalog storage device.

8. The system according to Claim 6, wherein said unnecessary word elimination unit extracts a word belonging to a given category and eliminates the word as an unnecessary word from said table if the word appears more frequently than a given standard in another category.

9. The system according to Claim 6, wherein said table broken down by category generated by said category table generation unit contains information on the word, a frequency of appearance of the word, and a part of speech of the word.

10. An unnecessary word determination method in a document automatic classification system, comprising the steps of:

extracting a word contained in a document for each category from a storage device storing a learning document set;

generating a list containing information on a frequency of appearance of the extracted word for each category;

recognizing a frequency of appearance in other categories of a given word belonging to a given category by using the generated list; and

determining an unnecessary word for each category on the basis of the recognized frequency of appearance.

11. The method according to Claim 10, wherein, in said step of determining the unnecessary word, the unnecessary word is determined according to whether one word selected from the given category appears in said other categories more frequently than a given standard.

12. The method according to Claim 11, wherein said given standard is a value obtained from the number of documents in said other categories and a predetermined given threshold.

13. The method according to Claim 11, wherein said given standard is determined according to said frequency of the word in said other categories and a total frequency of all words in said other categories.

14. An unnecessary word determination method in a document automatic classification system, comprising the steps of:

acquiring information on words for each category from a document set classified according to category stored in a storage device;

recognizing a frequency of appearance in other categories of a word belonging to a given category on the basis of the acquired information; and

determining whether the word is unnecessary for identifying the given category on the basis of the recognized frequency.

15. The method according to Claim 14, further comprising the steps of:

generating a document classification catalog by eliminating words determined to be an unnecessary word; and

storing said classification catalog into the storage device.

16. The method according to Claim 1, further comprising the step of performing classification processing for classification target documents by using the classification catalog stored in said storage device.